Russian Bridging Anaphora corpus

Anna Roitberg
IMPB RAS
Branch of KIAM RAS, Russia
HSE RSU, Russia
cvi@yandex.ru

Denis Khachko
IMPB RAS
Branch of KIAM RAS, Russia
mordol@lpm.org.ru

Abstract

In this paper, we present a bridging anaphora corpus for Russian, introduce a syntactic approach for bridging annotation and discuss the difference between the syntactic and semantic approaches. We also discuss some special aspects of bridging annotation for Russian and other languages where definite nominal groups are not marked so frequently as e.g. in Romance or Germanic languages. In the end we list the main cases of annotator disagreement.

1. Introduction

Anaphoric links are very important for text cohesion. In 1975, Clark (Clark, 1975) contrasted direct anaphora and indirect anaphora (bridging). The term *direct anaphora* (1) is used for cases where anaphorically linked entities are coreferent. In the case of an *indirect anaphora* (2), anaphorically linked entities are not coreferent, but associated reflecting more complicated semantic relations.

(1) *I looked at his car yesterday. A really old* ***vehicle***.

(2) *I looked at his car yesterday.* ***The door*** *was rusty.*

In researches of the direct anaphora, the terms *anaphoric element* (for vehicle in (1)) and *antecedent* (for car in (2)) are usually used. In the bridging anaphora researches, the terms *bridging element* (instead of anaphoric element) and *anchor* (instead of antecedent) are more common.

Bridging anaphora involves a very wide spectrum of semantic relations from the part-whole relations to relations between different arguments accompanying a single predicate. So all studies on bridging, as we know them, limit the number of bridging relations which they work with. The most common way to constrain the amount of bridging types is to consider several types of semantic relations, the most popular of which are part-whole and set-subset relations. This approach is used in inspiring the Poesio's projects: GNOME corpus (Poesio, 2000; Poesio et al., 2004) and ARRAU for English (Poesio and Artstein, 2008); in the second edition of the ARRAU corpus, the set of bridging relations became wider but it is still limited on the semantic ground. The same approach can be found in the CESS-ECCE corpus for Spanish (Recasens et al., 2007) and AnCora for (Recasens, 2008) Catalan, and PAROLE for French (Gardent et al., 2003). A wide spectrum of semantic types of bridging relations is annotated in the Prague Dependency Treebank for Czech (Mikulová et al., 2017; Nedoluzhko and Mírovskỳ, 2011). The semantic constraints are also used in recent researches: multilingual corpus for English, German and Russian (Grishina, 2016) and GUM corpus for English (Zeldes, 2017)

We call this approach *semantically oriented*.

The second approach appeared through development of computational methods in linguistics. No semantic constraints are used here. This approach is less popular, but it is used in (Hou et al., 2013) where very impressive results are shown. The state of the art system for bridging resolution (Hou et al., 2016) is based on this corpus.

The paper is structured as follows: Section 2 provides our syntactic oriented approach for bridging anaphora and introduces the term *genitive bridging*, Section 3 presents RuGenBridge corpus and annota-

Keywords: anaphora, bridging annotation, bridging corpus

tion scheme, Section 4 describes inter-annotator agreement, in Section 5 we discuss main cases of typical disagreement.

2. Bridging anaphora annotation approach for Russian

Bridging anaphora is a very complicated high-level phenomenon and the research is in its infancy. Due to this, there are no standard annotation schemes up to the present. The annotation scheme used usually corresponds appropriately to the study purposes.

Our main goal is to develop an automatic bridging recognition system (set of classifiers) based on machine learning techniques. First of all, we considered bridging relations between noun phrases (NP) and marked just heads of noun phrases. Recall that there are no articles in Russian, we could not focus solely on definite NPs as in studies for Romano-Germanic languages. Afterwards we decided to annotate only those features which could be useful for ongoing classifiers. That led to a decision not to use the semantic-oriented approach, because we cannot utilize and implement this knowledge. Russian WordNet and similar resources are not as developed as English analogues.

2.1. Genitive bridging

On that basis we concentrated on a new formal-oriented, syntactic approach. We decided to restrict the amount of bridging cases to one syntactic construction, more specifically, the genitive construction. The genitive construction *N+Ngen* is very common in Russian, it typically marks possessive relations (in a broad sense). So it is associated with (but not limited to) such semantic relations as *item – possessor*, *part – whole* etc.

Therefore, we annotated the cases of bridging anaphora where the bridging element and anchor could form a grammatical genitive construction as in the following:

(3) *Ja kupil telefon, no **knopki** okazalis' slishkom malen'kimi.*
*'I bought the phone, but the **buttons** turned out to be too small'*

On the one hand, in the example above, the words that mean "phone" and "buttons" are anaphorically linked: there are not just some buttons but specifically the buttons of the previously mentioned phone. On the other hand, in Russian the anchor "phone" and the bridging element "buttons" can form a grammatical genitive construction bridging *element + anchor.Gen*: *"knopki telefona.Gen" 'the buttons of the phone'*. We called this kind of bridging relations *"genitive bridging"*.

In our corpus we annotated only cases of genitive bridging.

3. Corpus RuGenBridge

Our corpus materials were short news texts from online news agencies. Short means 100 – 200 words. We chose such short texts due to the complexity of this phenomenon: annotators make more mistakes in long texts, because of the difficulty of keeping in mind discourse relations over a large distance.

At the time of writing, we annotated 339 texts or 61076 tokens, and tagged 609 genitive bridging pairs.

All bridging cases were manually annotated using the BRAT tool[1]. Parts of speech and syntactic links were annotated automatically, using FreeLing[2] and MaltParcer[3] (Nivre et al., 2006), correspondingly.

On the engineering side, our corpus is a SQL database, which consists of 3 main tables: 1) Table of texts; 2) Tables of lemmas and 3) Tables of relations

3.1. Boundary markables

We postulate a principle of minimum possible size for markables. Where possible, we mark single nouns – the heads of the corresponding noun phrase. In "the smallest house in the lane" only a "house" will be

[1] `http://brat.nlplab.org`
[2] `http://nlp.lsi.upc.edu/freeling/`
[3] `www.maltparser.org`

marked. In the case of having an anchor (or a bridging element) as a named entity, we mark all the entity, so in *"Cherry Tree Lane"* we mark *"Cherry Tree Lane"*; the same for names of persons, organizations, geographic names etc.

3.2. Semantic labels

Despite not using the semantic approach to corpus annotation, we can still use some semantic labels for anchors and bridging elements to mark the most popular semantic types which could be relevant for future work.

The set of labels is given below.

1. Geo – for proper names of geographic objects (Brazil, Indian Ocean, Grand Canyon). Compare (4) and (5):

   (4) *The government of* ***Moscow.Geo*** *is continuing to discuss* ***transportation***.

   (5) *The government of* ***the city*** *is continuing to discuss* ***transportation***.

2. ORG – for proper and common names, refers to official organisations, public institutions etc.: government, Russian Orthodox Church, BBC. We take into account the contextual meaning of a noun phrase. Compare (6) and (7):

   (6) *BBC World Service.ORG has announced the extension of* ***the agreement****...*

   (7) *She used to listen to the BBC especially* ***news programs****...*

3. POST – job titles: president, coach, cardinal, priest, dean

   (8) *FC Barcelona.ORG has announced that* ***the coach.POST****was dismissed.*

The total number of semantic labels in RuGenBridge corpus is shown in Table 1.

Semantic label	Anchor	Bridging-element	Total
GEO	148	9	157
ORG	11	24	35
POST	22	-	22

Table 1: Semantic labels in RuGenBridge

These types of semantic labels were chosen in view of the fact that the lists of such lexical groups can be extracted from dictionaries and ontologies. This information can be used to construct a bridging anaphora recognition and resolution system, which was the main purpose of the project.

3.3. Bridging relations in RuGenBridge

Considering that we are using a new approach to bridging, we tried to analyze what types of bridging pairs (on semantic point of view) were annotated. We compared what kind of bridging relations become annotated by using each of the approaches. As a reference, we choose semantically oriented annotation scheme, using in Prague Dependency Treebank (PDT) – (Nedoluzhko and Mírovskỳ, 2011). There are two advantages of this scheme for our project: 1) it is the one of the most developed semantic oriented scheme, 2) it was constructed for Czech – Russian's relative language. There are 6 types of bridging relations are emphasized in PDT: (1) PART-WHOLE and WHOLE-PART, as e.g. in face – eyes), (2) SUBSET- SET and SET-SUBSET, as in a group of students – some students – a student), (3) the relation between an entity and a singular function on this entity (subtypes P-FUNCT and FUNCT-P, as in company – director) (4) the relation between coherence-relevant discourse opposites (type CONTRAST,

as in black flags – white flags), (5) non-coreferential explicit anaphoric relation (type ANAPH, as in first world war – at that time) and (6) further underspecified group REST consisting of six other bridging subtypes (e.g. relations between family members, event – argument, locality – inhabitant, etc.). We provided two experiments, fully described in (Roitberg and Nedoluzhko, 2016). In the first experiment, eight texts of the corpora were annotated with both schemes: genitive bridging and PDT. We annotated 69 bridging pairs using the PDT scheme, 22 pairs using the RuGenBridge scheme, but there were only 7 coincidence cases. During the second experiment, we added PDT annotation marks (for semantic type of bridging relations) for all genitive bridging pairs in 200 texts (more than a half of texts). All bridging relations using in PDT are listed in (Nedoluzhko et al., 2009). We analyzed what semantic types are most frequent among the genitive bridging pairs. The most frequent were: PART-WHOLE (WHOLE-PART), SET-SUB (SUB-SET), FUNCT-P (P-FUNCT); the last type is often used for government positions (parliament – speaker). Besides bridging relations we annotate coreference chains, but only for entities that were previously annotated as anchors or bridging elements.

We also analyzed which types of bridging relations were annotated with the genitive bridging approach are usually missed when semantic approach to annotation is used. We found out that just a half of genitive bridging pairs can be marked with any of semantic PDT labels. There are two main groups of cases, which cannot be classified as any of of PDT types of bridging: 1) the pairs that reflect text cohesion more than semantic relations. For example geographic names – something located there, like 'Moscow – hospitals'; and 2) bridging relations between non-referential nouns, like 'oil – barrel'; non-referential nouns were not marked in PDT on formal ground. Such syntactic oriented approach can be useful for those researches who study these types of bridging anaphora.

4. Evaluating the quality

4.1. Inter-annotator agreement

High-level annotation is a challenge. The higher-level phenomenon is less strictly described in theoretical models, so there are a lot of borderline cases which are difficult to annotate. Moreover, the discourse annotation requires close attention because an annotator has to keep in mind the text as a whole, not just a solitary word. This said, the inter-annotator agreement in high-level annotation is usually not as high as, for example, in part of speech tagging.

Corpus RuGenBridge was annotated by three annotators and a supervisor. The statistics for all annotations are shown in Table 2.

	Annotator 1	**Annotator 2**	**Annotator 3**
Anchors	167	419	663
Bridging elements	273	620	846
Bridging links	273	620	846

Table 2: Labels statistics for different annotations

The first annotator was inclined to miss some genitive bridging cases, whereas in contrast other annotators (especially Annotator 3) marked several false pairs.

In spite of visible differences, the level of agreement (F-measure) between Annotator 1 and Annotator 2 was sufficient in more detail see (Table 3).

An 1. Total links	**An 2. Total links**	**True positive**	**False positive**
273	620	147	473

Table 3: Inter-annotator agreement between Annotator 1 and Annotator 2.

While computing the Inter-annotator agreement, we considered one annotation as a gold standard and computed F-measure regarding this annotation.

An 1. Total links	An 3. Total links	True positive	False positive
273	846	105	741

Table 4: Inter-annotator agreement between Annotator 1 and Annotator 3.

We used F-measure for inter-annotator agreement in line with (Nedoluzhko and Mírovskỳ, 2013). The more widespread Cohen's kappa can not be applied to such rare phenomenon as bridging anaphora. For rare phenomenon the number of no-no cases (close to *true negatives*) in confusion matrix is incomparably higher than the number of yes-yes cases (close to *true positives*) and yes-no cases, so Cohen's kappa would always be in the neighborhood of 1.

As presented in Table 3, we considered Annotation 2 regarding Annotation 1. Notice that *True positive* – is the set of bridging pairs that are matched between Annotator 1 and Annotator 2; *true negative* – is the set of pairs which were labeled as bridging by Annotator 2 and in contrast were not labeled as bridging by Annotator 1.

On account of the data represented in Table 3, the inter-annotator agreement between Annotator 1 and Annotator 2 is at F-measure = 0.71

Unfortunately the level of agreement between Annotator 3 and other annotators was unacceptably low as shown in Table 4.

The F-measure for this pair of annotations is just F=0.37. Since this annotation contained multiple errors, we did not use this annotation in in our results.

In the final release of the RuGenBridge Corpus the supervisor combined the annotations of Annotator 1 and Annotator 2 and removed all false pairs, which in truth were not the cases of genitive bridging.

4.2. Cases of typical disagreement

Bridging annotation requires both solid annotator's experience and well-thought-out guidelines, but the main problem for annotators is to keep in mind the text and to concentrate on deciding if the noun in question has a bridging link to some anchor.

We summarized up the main types of inter-annotator agreement errors. In obvious way, there are three main groups of errors: 1) to omit a bridging pair, 2) to add a false pair, 3) to choose the wrong anchor for some bridging element. Beside errors, there are also some cases of insignificant differences between annotations.

We provide examples of each case in what follows.

4.2.1. Omitting of bridging-pairs

Omission of bridging-pairs is obviously the most common type of annotation errors, but happens more frequently where a bridging element and an anchor are linearly close to each other. The anaphoric link seems to be trivial in such cases, but it should be annotated on formal ground.

(9) *Prezident v* ***obrash'enii*** *zayavil...*
'The President announced in the ***address****...'*

It is worth mentioning that to miss bridging pairs, to miss bridging pairs at a long-distance (those with an anchor in the very beginning of the text and bridging element at the end of the text) bridging relations was the second most frequent type of errors of this sort.

4.2.2. Adding false pairs

Genitive bridging criteria is formal and "machine-friendly", but in some situations it was difficult to follow this criteria, because there were some cases semantically close to genitive bridging relations. Sometimes such pairs were annotated by mistake. One of the most frequent cases was bridging relations between two geographic objects, where one is a part of another. In Russian, two geographic names can usually form grammatical genitive construction, when the head is a name of a country and the dependence is a name of some region of the country. In Russian, the dependence usually contains such general words

as *oblast', kraj* means *'region'*. Several expressions of that type can be used in genitive constructions as follows:

(Part_Geo) + (Whole_Geo).Gen

"Moskovskaja oblast" and "Rossijskaja Federacija" can form a grammatical genitive construction, see Example (10).

(10) *Moskovskaja oblast' Rossijskoj Federacii_Gen* 'Moscow region of Russia federation'

However, even more expressions can not form grammatical genitive construction on formal ground even though they are semantically very close to previous ones. Example (11) below is ungrammatical.

(11) **Sibir' Rossijskoj Federacii_Gen*

4.2.3. Mismatches in coreference chains

In the RuGenBridge corpus annotation guideline it was mentioned that the annotator should choose the linearly closest preceding anchor. In several cases, the annotators missed the closest anchor and made a link to some other coreferential NP. We consider cases of this sort as insignificant, so such errors was ignored while computing inter-annotator agreement.

(12) *(. . .)* Tol'jatiazota, v sluchae esli ne soglashus' na ih uslovija po prodazhe predpriajatija (. . .) ih tsel' rejderskij zahvat predprijatija (. . .) ne lehche li bylo by vykupit' dolyu **minoritarijev**.
'Of"Tol'jatiazot", If I do not accept their conditions for a business transfer (. . .) their goal is a asset-grabbing (. . .) was not it easy to buy out the (. . .) ***minority interest'***

One of Annotators drew an arrow from "minority interest" to "business" and the second annotator connected the "minority interest" to "Tol'jatiazot" (the name of the company). "Tol'jatiazot" and "business"are coreferentail expressions.

4.2.4. Comprehension disagreement errors

A minor proportion of errors was caused by different comprehension of texts as in Example bellow.

(13) *V Instagrame Papy Rimskogo pojavilas' fotografija Papy, obnimajush'ego dvuh devochek s sindromom Dauna s zhelto-goluboj lentoj v* ***rukah*** *.*
'In Papa's Instagram a photo appeared of Papa, hugging two girls with Down syndrome, holding yellow and blue ribbons in the ***hands****.'*

One annotator linked the bridging element "hands" with anchor "Papa", while the other annotator connected "hands" with "girls".

Importantly, in Russian a possessive pronoun before "hands" is not required, so there is a case of ambiguity.

It is interesting to note, that our automatic bridging recognition system marked highly likely both mentioned variants as bridging relations.

5. Conclusion

We have described a syntax-oriented annotation scheme used in the RuGenBridge corpus. The RuGenBridge corpus represents an inventory of bridging anaphora relations which are not limited to common semantic relations such as part-whole, set-subset etc. We have also shared an experience in bridging anaphora annotation. In line with our expectations, the development of the corpus reveals the complexity of discourse-level annotation, which leads to a lower level of inter-annotator agreement. To increase inter-annotator agreement, we consider training future annotators in discourse theory in general and especially in anaphora theory.

The RuGenBridge corpus can be used as a training and test data set for bridging anaphora recognition; see our pilot results in (Roitberg and Khachko, 2017). The corpus is available on request. The

supplementary materials on the project are available on GitHub repository [4].

References

Clark, H. H. (1975). Bridging. In *Proceedings of the 1975 workshop on Theoretical issues in natural language processing*, pages 169–174. Association for Computational Linguistics.

Gardent, C., Manuélian, H., and Kow, E. (2003). Which bridges for bridging definite descriptions. In *Proceedings of the EACL 2003 Workshop on Linguistically Interpreted Corpora*, pages 69–76.

Grishina, Y. (2016). Experiments on bridging across languages and genres. In *Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016)*, pages 7–15.

Hou, Y., Markert, K., and Strube, M. (2013). Global Inference for Bridging Anaphora Resolution. In *HLT-NAACL*, pages 907–917.

Hou, Y., Markert, K., and Strube, M. (2016). Unrestricted Bridging Resolution. *Computational Linguistics*, (Just Accepted):1–68.

Mikulová, M., Mírovskỳ, J., Nedoluzhko, A., Pajas, P., Štěpánek, J., and Hajič, J. (2017). PDTSC 2.0-Spoken Corpus with Rich Multi-layer Structural Annotation. In *International Conference on Text, Speech, and Dialogue*, pages 129–137. Springer.

Nedoluzhko, A. and Mírovskỳ, J. (2011). Annotating extended textual coreference and bridging relations in the Prague Dependency Treebank. *Annotation manual. Technical report*, (44).

Nedoluzhko, A. and Mírovskỳ, J. (2013). Annotators' Certainty and Disagreements in Coreference and Bridging Annotation in Prague Dependency Treebank. In *Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013)*, pages 236–243.

Nedoluzhko, A., Mírovskỳ, J., Ocelák, R., and Pergler, J. (2009). Extended coreferential relations and bridging anaphora in the Prague Dependency Treebank. In *Proceedings of the 7th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2009), Goa, India*, pages 1–16.

Nivre, J., Hall, J., and Nilsson, J. (2006). Maltparser: A data-driven parser-generator for dependency parsing. In *Proceedings of LREC*, volume 6, pages 2216–2219.

Poesio, M. and Artstein, R. (2008). Anaphoric Annotation in the ARRAU Corpus. In *LREC*.

Poesio, M., Mehta, R., Maroudas, A., and Hitzeman, J. (2004). Learning to resolve bridging references. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 143. Association for Computational Linguistics.

Poesio, M. (2000). Annotating a Corpus to Develop and Evaluate Discourse Entity Realization Algorithms: Issues and Preliminary Results. In *LREC*.

Recasens, M., Martı, M. A., and Taulé, M. (2007). Text as scene: Discourse deixis and bridging relations. *Procesamiento del lenguaje natural*, 39:205–212.

Recasens, M. (2008). Discourse deixis and coreference: Evidence from AnCora.

Roitberg, A. and Khachko, D. (2017). Bridging Anaphora Resolution for the Russian Language. In *Proceeding of 23rd Conference on Computational Linguistics and Intellectual Technologies Dialogue-2017*.

Roitberg, A. and Nedoluzhko, A. (2016). Bridging Corpus for Russian in comparison with Czech. In *Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016)*, pages 59–66.

Zeldes, A. (2017). The GUM corpus: creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

[4] `https://github.com/Anna-Roitberg/RuGenBridge`